IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 10, October 2017

Study & Implementation on Speech Synthesis by Sparse Modelling & Error Minimization in Noisy Environment

Gagandeep Kaur[1], Mr. Naveen Sharma[2]

Computer Science & Engineering Deptt. ICL Group of Colleges, Ambala, India[1,2]

Abstract: The aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. In this work, it provides speech enhancement under different noisy conditions. The use of speech enhancement algorithm removes or reduces the presence of noise. The aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it presents a method for speech enhancement using mask estimation iteratively. In this work, it provides the concept of optimization of cost function by iterative method. This helps for reducing the noise from signal. All simulations are implemented in MATLAB.

Keywords: Speech Enhancement, Speech Processing, Noise Filtering, Sparse Representation etc.

I. INTRODUCTION

During conversation, both hearing and speaking adapt to the background noise in a noisy environment. It is therefore possible to have a conversation in quite disturbing background noise environments. However, when the conversation takes place over the telephone disturbances are more annoying. The disturbances are a problem since the brain will not get the extra visual and other background information when interpreting the speech. The speech signal transmitted to the other party is picked up by a microphone connected to the telephone. The microphone signal contains both speech and noise at some ratio (Speech to Noise Ratio, SNR) depending on, for example, how far the microphone is mounted from the speaker's mouth.

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. In practice, a convolutive noise should be rather considered due to the reverberation. However, it is usually assumed that the noise is additive since it makes the problem simpler and also the developed algorithms based on this assumption lead to satisfactory results in practice. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. There are various applications of speech enhancement in our daily life.

For example, consider a mobile communication where you are located in a noisy environment, e.g., a street or inside a car. Here, a noise reduction approach can be used to make the communication easier by reducing the interfering noise. A similar approach can be used in communications over internet, such as Skype or Google Talk. Speech enhancement algorithms can be also used to design robust speech/speaker recognition systems by reducing the mismatch between the training and testing stages. In this case, a speech enhancement approach is applied to reduce the noise before extracting a set of features.

In literature, some proposed a novel multi-channel speech enhancement method by combining the wiener filtering and subspace filtering with a convex combinational coefficient. It investigated a multi-channel de-noising auto-encoder (DAE)-based speech enhancement approach. In recent years, deep neural network (DNN)-based monaural speech enhancement and robust automatic speech recognition (ASR) approaches have attracted much attention due to their high performance. It also proposed a sparse hidden Markov model (HMM) based single-channel speech enhancement method that models the speech and noise gains accurately in non-stationary noise environments. Some presented a harmonic phase estimation method relying on fundamental frequency and signal-to-noise ratio (SNR) information estimated from noisy speech. The proposed method relies on SNR-based time-frequency smoothing of the unwrapped phase obtained from the decomposition of the noisy phase.

The paper is ordered as follows. In section II, it represents introduction of speech signals. In Section III, It defines the description of speech level estimation system. Section IV describes the problem definition of system. Finally, conclusion is explained in Section V.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 10, October 2017

II. SPEECH SIGNALS

A speech signal consists of three classes of sounds. They are voiced, fricative and plosive sounds. Voiced sounds are caused by excitation of the vocal tract with quasi-periodic pulses of airflow. Fricative sounds are formed by constricting the vocal tract and passing air through it, causing turbulence those results in a noise-like sound. Plosive sounds are created by closing up the vocal tract, building up air behind it then suddenly releasing it.
Figure 1 shows a discrete time representation of a speech signal. By looking at it as a whole we can tell that it is non-stationary. That is, its mean values vary with time and cannot be predicted using the above mathematical models for random processes. However, a speech signal can be considered as a linear composite of the above three classes of sound, each of these sounds are stationary and remain fairly constant over intervals of the order of 30 to 40 ms [2].
Speech sounds can be broadly divided into two categories: voiced and unvoiced. Voiced sounds are produced when the vocal folds are vibrating, producing a quasi periodic signal, while unvoiced sounds are articulated without vibration of the vocal folds. Speech consists of a sequence of vowels and consonants together with brief silences between phonemes and words. Vowels are created by a voiced sound without any constriction in the vocal tract.

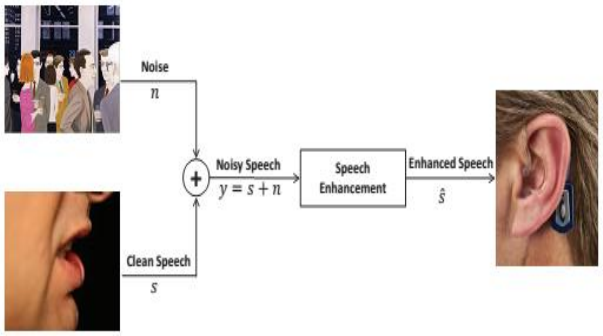

Figure 1: Speech Enhancement System with Corrupted Noise [1]

Estimation of a clean speech signal from a noisy recording is a typical signal estimation task. But due to the non-stationary of the speech and most of the practical noise signals, and also due to the importance of the problem, significant amount of research has been devoted to this challenging task. Single-channel speech enhancement algorithms e.g. use the temporal and spectral information of speech and noise to design the estimator. In this case, only the noisy recording obtained from a single microphone is given while the noise type, speaker identity or speaker gender is usually not known. Multichannel or multi microphone noise reduction systems, utilize the temporal and spectral information as well as the spatial information to estimate a desired speech signal from the given noisy recordings Consonants, however, can be originated by a voiced or an unvoiced sound and are classified as:

- **Stops:** which occur when the air flow is blocked and suddenly released
- **Nasals:** produced when the air is stopped in the oral cavity but not through the nasal cavity.
- **Approximants:** produced when there is a constriction but not narrow enough to result in turbulence.
- **Fricatives:** a narrow constriction in the vocal tract resulting in a turbulent air flow.

To illustrate the passage of a speech signal from talker to listener, a typical single channel speech recording chain is shown in Fig. 2. The desired speech signal passes through a convolutive acoustic channel before reaching the microphone, where it is combined with sound from other acoustic sources in the environment and it is transduced into the electronic domain. The speech signal can become degraded by further additive noise as well as by possible non-linear distortion within the electronic domain.

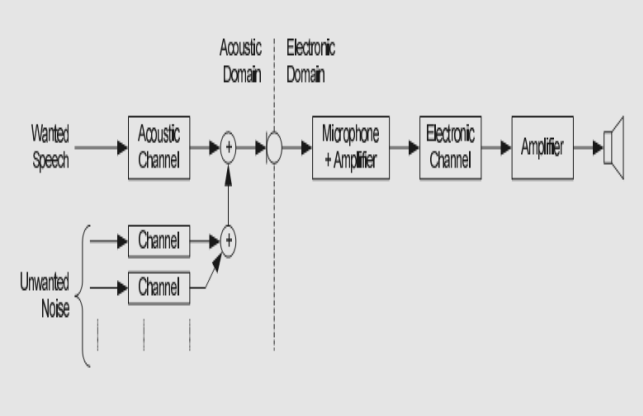

Figure 2: Typical Speech Recording Chain [5]

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 10, October 2017

It is convenient to classify speech signal degradations into the following three classes which differ in their causes and potential remedies:

- Additive background noise that can arise in either the electronic or acoustic domains, although serious signal degradation is normally caused only by acoustic noise from unwanted sources in the environment;
- Convolutive effects including echo and reverberation; and
- Non-linear speech distortion which may, for example, be introduced by amplitude limiting or clipping in the microphone, amplifier or Coder-Decoder (CODEC).

In recent decades a diverse range of solutions has been proposed to address these degradation effects. Speech enhancement techniques aim to restore corrupted speech signals by removing or compensating for degradation without damaging the speech signal itself. The work in this thesis is concerned with the enhancement of single channel speech signals that have been corrupted by levels of additive noise that are high enough to affect the intelligibility of the speech.

III. DESCRIPTION OF SPEECH LEVEL ESTIMATION

In general, speech enhancement methods can be categorized into two broad classes: unsupervised and supervised. In unsupervised methods such as Wiener and Kalman filters and estimators of the speech DFT coefficients using super-Gaussian priors, a statistical model is assumed for each of the speech and noise signals, and the clean speech is estimated from the noisy observations without any prior information on the noise type or speaker identity. Speech consists of a sequence of vowels and consonants together with brief silences between phonemes and words [10]. Vowels are created by a voiced sound without any constriction in the vocal tract. Noise, in contrast to speech, can originate from any kind of source and have any spectral and temporal characteristics. There are, however, some common assumptions made about the noise when approaching the speech enhancement problem:
(i) The power spectrum of noise is more stationary than that of speech, and
(ii) Speech and noise are statistically independent.
It considers a multi-path environment where one source and two sensors are presented; the two sensors are located at different distances from the same source. The received signal at the two microphones can be modelled as:

$r_1(t) = s(t) + n_1(t), \qquad 0 \le t \le T \qquad r_2(t) = s(t - D) + n_2(t)$

Where $r_1(t)$ and $r_2(t)$ are the outputs of the two microphones that are separated spatially, $s(t)$ is the source signal, $n_1(t)$ and $n_2(t)$ are representing the additive noises.'T', the observation interval ,and 'D',the time delay between the two received signals. The signal and noises are assumed to be uncorrelated having zero-mean and Gaussian distribution. Our objective is to estimate this 'D' and thus the problem 'Time Delay Estimation'.

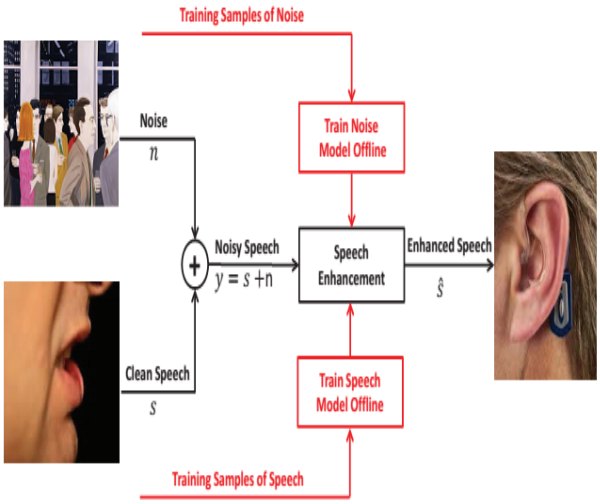

Figure 3: Supervised Speech Enhancement System

The majority of the energy in a speech signal is concentrated in the voiced intervals. In the time-frequency domain, most of the voiced speech energy is located in a small number of harmonic peaks that remain detectable even at poor SNRs. In this section, we propose a method to estimate the speech active level at low SNRs from the energy of the harmonic peaks during voiced intervals. Intelligibility and pleasantness are difficult to measure by any mathematical algorithm. Usually listening tests are employed.
However, since arranging listening tests may be expensive, it has been widely studied how to predict the results of listening tests. The central methods for enhancing speech are the removal of background noise, echo suppression and the process of artificially bringing certain frequencies into the speech signal. First of all, every speech measurement

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 10, October 2017

performed in a natural environment contains some amount of echo. Echoless speech, measured in a special anechoic room, sounds dry and dull to human ear. In most cases the background random noise is added with the desired speech signal and forms an additive mixture which is picked up by microphone. It can be stationary or non stationary, white or colored and having no correlation with desired speech signal.
If the background noise is evolving more slowly than the speech, i.e., if the noise is more stationary than the speech, it is easy to estimate the noise during the pauses in speech. Finding the pauses in speech is based on checking how close the estimate of the background noise is to the signal in the current window. Voiced sections can be located by estimating the fundamental frequency. Both methods easily fail on unstressed unvoiced or short phonemes, taking them as background noise. On the other hand, this is not very dangerous because the effect of these faint phonemes on the background noise estimate is not that critical.
The algorithm provides a fundamental frequency estimate at every time-frame, together with a probability of each time-frame containing voiced speech. Identifying time-frames which contain sibilant phones is important for the preservation of periodic speech energy at high frequencies. Furthermore, an estimation of the power spectrum of the sibilant phone would also help identifying the frequency bands containing most of the sibilant speech energy.

IV. RESULTS & DISCUSSION

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise. The noise can be additive or convolutive. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. The main focus is to improve the cost of system.
In this work, it investigates a method without any pre-training of noise models. The only assumption about the noise is that it is different from the involved speech. Therefore, the noise estimation turns out to be finding the components which cannot be adequately represented by a well defined speech model. Given the good performance of deep learning in signal representation, a deep auto encoder (DAE) is employed for accurately modelling clean speech spectrum. Active noise suppression is a method in which the idea is to produce anti-noise into the listener's ear to cancel the noise. The delay must be kept very small to avoid producing more noise instead of cancelling the existing noise. In this section, we propose a method to estimate the speech active level at low SNRs from the energy of the harmonic peaks during voiced intervals. Both speech enhancement methods aimed at suppressing the background noise are (naturally) based in one way or the other on the estimation of the background noise. If the background noise is evolving more slowly than the speech, i.e., if the noise is more stationary than the speech, it is easy to estimate the noise during the pauses in speech. Finding the pauses in speech is based on checking how close the estimate of the background noise is to the signal in the current window. Voiced sections can be located by estimating the fundamental frequency. Both methods easily fail on unstressed unvoiced or short phonemes, taking them as background noise.

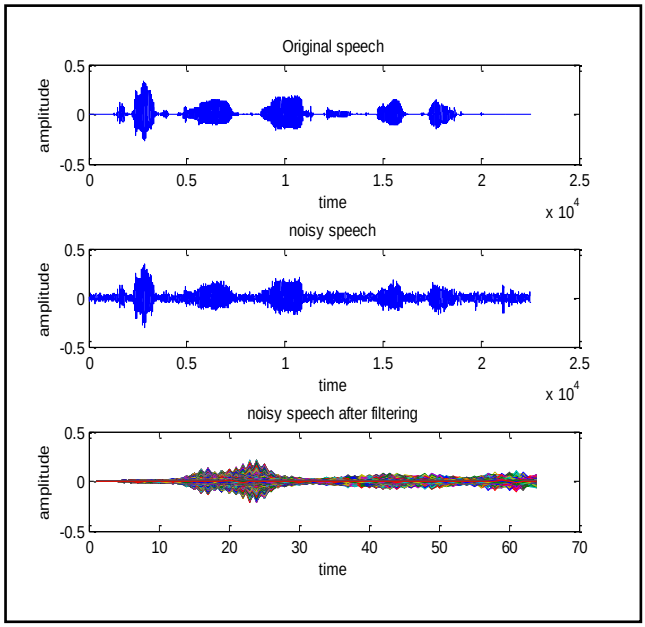

Figure 4: Original & Noisy Speech

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 10, October 2017

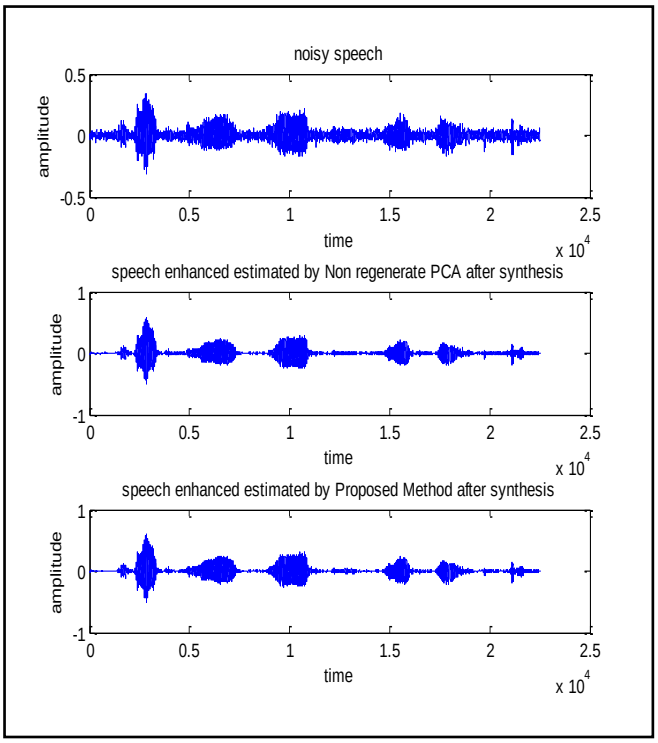

Figure 5: Speech Enhancement by Actual & Proposed Method

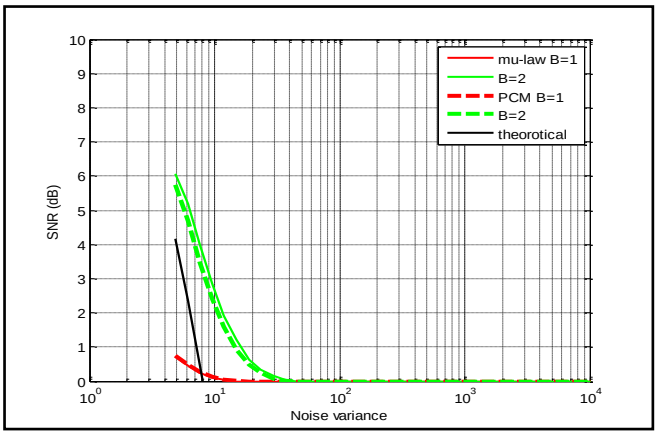

Figure 6: Proposed SNR Performance of System

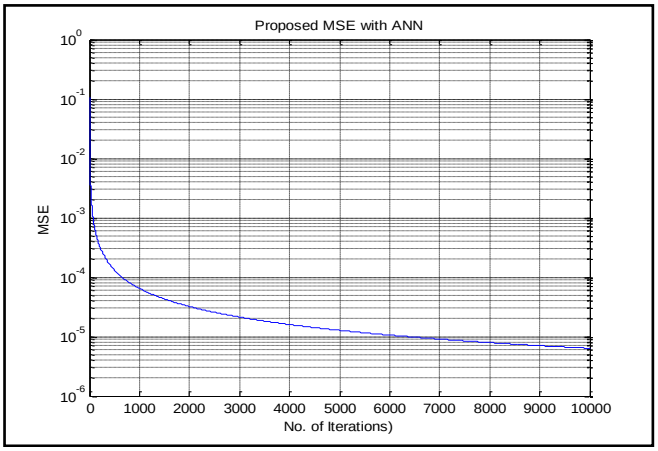

Figure 7: Proposed MSE Response

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 10, October 2017

The SNR results are shown in fig 6 & 7. The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It's called the mean squared error as you're finding the average of a set of errors.

V. CONCLUSION

Speech signals can be degraded in many ways during their acquisition in noisy environments and they can also be further degraded in the electronic domain. Serious signal degradation, however, is most commonly caused by noise from unwanted acoustic sources in the environment, which may affect the speech quality and/or intelligibility of the wanted signal. The noise can be additive or convolutive. Even this additive noise can reduce the quality and intelligibility of the speech signal considerably. Therefore, the aim of the noise reduction algorithms is to estimate the clean speech signal from the noisy recordings in order to improve the quality and intelligibility of the enhanced signal. Due to this, it proposes a method for speech enhancement using mask estimation iteratively. In this work, it provides speech enhancement under different noisy conditions. After this, it provides the performance comparison with non regenerative method in terms of cost and rank of matrix.
In future, the development of new ones to extract more speech information or on the enhancement of the mask estimate. Future work to improve the algorithm could include the application of temporal continuity constraints to the voicing probability estimate.

REFFERENCES

[1] Meng Sun, Xiongwei Zhang, Hugo Van hamme, "Unseen Noise Estimation Using Separable Deep Auto Encoder for Speech Enhancement", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 24, No. 1, January 2016.
[2] Shoko Arakit, Tomoki Hayashi, "Exploring Multi-Channel Features for Denoising-Auto-encoder-Based Speech Enhancement", IEEE 2015.
[3] Zheng Gong and Youshen Xia, "Two Speech Enhancement-Based Hearing Aid Systems and Comparative Study", IEEE International Conference on Information Science and Technology, April 24-26, 2015.
[4] Feng Deng, Changchun Bao, "Sparse Hidden Markov Models for Speech Enhancement in Non-Stationary Noise Environments", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 11, November 2015.
[5] Pejman Mowlaee and Josef Kulmer, "Harmonic Phase Estimation in Single-Channel Speech Enhancement Using Phase Decomposition and SNR Information", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 9, September 2015.
[6] Swati R. Pawar, Hemant kumar B. Mali, "Implementation of Binary Masking Technique for Hearing Aid Application", IEEE International Conference on Pervasive Computing, 2015.
[7] Xia Yousheng, Huang Jianwen, "Speech Enhancement Based on Combination of Wiener Filter and Subspace Filter", IEEE 2014.
[8] Zhang Jie, Xiaoqun Zhao, Jingyun Xu, "Suitability of Speech Quality Evaluation Measures in Speech Enhancement", IEEE 2014.
[9] Atsunori Ogawa, Keisuke Kinoshita, Takaaki Hori, "Fast Segment Search For Corpus-Based Speech Enhancement Based On Speech Recognition Technology", IEEE International Conference on Acoustic, Speech and Signal Processing, 2014.
[10] AN.SaiPrasanna, Iyer Chandrashekaran, "Real Time Codebook Based Speech Enhancement with GPUs", International Conference on Parallel, Distributed and Grid Computing, 2014.
[11] Zawar Shah, Ather Suleman, Imdad Ullah, "Effect of Transmission Opportunity and Frame Aggregation on VoIP Capacity over IEEE 802.11n WLANs", IEEE 2014.
[12] Lee Ngee Tan, Abeer Alwan, "Feature Enhancement Using Sparse Reference And Estimated Soft-Mask Exemplar-Pairs For Noisy Speech Recognition", IEEE International Conference on Acoustic, Speech and Signal Processing, 2014.
[13] Seung Yun, Young-Jik Lee, and Sang-Hun Kim, "Multilingual Speech-to-Speech Translation System for Mobile Consumer Devices", IEEE Transactions on Consumer Electronics, Vol. 60, No. 3, August 2014.
[14] Christian D. Sigg, Tomas Dikk, "Speech Enhancement Using Generative Dictionary Learning", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 6, August 2012.
[15] H. Veisi H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement", IET Signal Processing, 2012.